Journal of Economic Theory **86**, 148–156 (1999)
Article ID jeth.1999.2507, available online at http://www.idealibrary.com on IDEAL®

NOTES, COMMENTS, AND LETTERS TO THE EDITOR

Can Pre-arranged Matches Be Avoided in Two-Sided Matching Markets?

Tayfun Sönmez*

Department of Economics, University of Michigan, Ann Arbor, Michigan 48109-1220 and College of Economics and Administrative Sciences, Koç University, 80860 Istanbul, Turkey

Received January 23, 1998; revised December 29, 1998

We study *manipulation via pre-arranged matches* in the context of centralized two-sided matching markets. We show that the solution that is used to match the hospitals and medical residents in the United States, namely the hospital-optimal stable rule, is manipulable in this way. Unfortunately this is a general problem: We show that there is no solution that is both stable and non-manipulable. *Journal of Economic Literature* Classification Numbers: C71, C78, D71, D78.

1. INTRODUCTION

In this paper we study the vulnerability of centralized *two-sided matching markets* (Gale and Shapley [1]) to *manipulation via pre-arranged matches*. There are two finite and disjoint sets of agents, say medical interns and hospitals. Each hospital has a capacity which is the maximum number of interns it can employ. Each intern has a preference relation over the set of hospitals and being unemployed, and each hospital has a preference relation over the set of groups of interns. An allocation is a matching of interns and hospitals such that no hospital is assigned more interns than its

* I am grateful to Steven Ching, Joseph Greenberg, Alvin Roth, William Thomson, an anonymous referee, and an anonymous associate editor for their extensive comments. I thank Ted Bergstrom, Peter Morgan, Elliot Peranson, Jim Schummer, Ennio Stacchetti, Yao-Chou Tsai as well as the seminar participants at the McGill University, University of Michigan, University of Pittsburgh, ESEM 1996 (Istanbul), ESSET 1997 (Gerzensee), the seventh Stony Brook Summer Festival on Game Theory, the conference Axiomatics of Resource Allocation (Montréal 1996), the NATO ASI workshop Game Theory and Resource Allocation: The Axiomatic Approach (Stony Brook 1997) for their useful suggestions. I gratefully acknowledge the research support of the National Science Foundation via Grant SBR-9709138. All errors are my own responsibility.

0022-0531/99 $30.00

capacity and no intern is assigned more than one hospital. A matching is *stable* if (i) no hospital prefers keeping a position vacant rather than filling it with one of its assignments, (ii) no intern prefers remaining unemployed to his/her assignment, and (iii) there is no unmatched hospital-intern pair such that the intern prefers the hospital to his/her assignment and the hospital prefers the intern either to one of its assignments or to keep a vacant position (in case it has one). Among the stable matchings, there is one which is preferred to any stable matching by all the hospitals (Roth [4]). This matching is called the *hospital-optimal stable matching*. Not surprisingly, this matching is the worst stable matching for all the interns. There is an analogous stable matching that favors the interns and it is called the *intern-optimal stable matching*.

The stability criterion has been central to studies concerning two-sided matching markets. (See Roth and Sotomayor [10] for an extensive analysis.) Indeed Roth [4] shows that the National Resident Matching Program (NRMP) has been using the *hospital-optimal stable rule* (the solution that selects the hospital-optimal stable matching) to match medical interns and hospitals in United States since 1950.[1] This solution has been quite successful after years of chaos in the U.S. hospital-intern market. It is widely agreed that the stability of this solution is the key factor in its relative success. For example in the United Kingdom several unstable solutions are abandoned due to various problems. (See Roth [7].) One of the problems that is widely observed in markets where unstable solutions are adopted is that agents pre-arrange matches before the formal procedure and thus circumvent it. This is expected in the case of unstable solutions. However, such behavior is also observed in the U.S. hospital-intern market and Canadian lawyer markets in spite of the use of stable solutions. Roth and Xing [11] offer a number of explanations for this phenomenon that are based on a multi-period model in which there are some uncertainties about interns' qualifications in early periods.[2] Roth [6] on the other hand states, "Unstable matchings give some market participants incentives to circumvent the formal procedures. Open questions remain concerning..., and whether there may be environments in which even stable procedures are prone to unraval." Motivated by this question we approach this phenomenon from a manipulation point of view and ask the following questions: Do the stable solutions that are used in these markets (for example, the hospital-optimal stable rule) give agents the incentives to pre-arrange their matches? If the answer is yes, then to what extent can we avoid this problem via adopting other stable solutions?

[1] Strictly speaking a variant of this rule is used to minimize the complications associated with married couples and positions that are linked. See Roth [8] for a description of this rule.

[2] See also Li and Rosen [2] for a similar approach in assignment problems.

It turns out that the hospital-optimal stable rule is *manipulable via pre-arrangement of matches.* Furthermore there is no solution that is both stable and non-manipulable via pre-arrangement of matches. We next show that manipulation via pre-arranged matches can happen in two ways: (i) *Type*-1 *manipulation.* A pre-arranged match between a hospital and an intern who would be assigned to the hospital in the absence of manipulation can strictly benefit the hospital (without hurting the intern.) (ii) *Type*-2 *manipulation.* A pre-arranged match between a hospital and an intern that it finds strictly worse than any intern who would be assigned to the hospital in the absence of manipulation, can strictly benefit both the hospital and the intern.

2. THE MODEL

A (*many-to-one*) *matching problem* is a four-tuple (H, I, q, R). The first two components are sets of hospitals $H=\{h_1, ..., h_n\}$ and interns $I=\{i_1, ..., i_m\}$. We will denote the set of all agents $H \cup I$ by N whenever we can so as to simplify the notation. The third component is a vector of non-negative natural numbers $q=(q_{h_1}, ..., q_{h_n})$, where $q_{h_i} \geqslant 0$ is the capacity of hospital $h_i \in H$.[3] Let q_{-h} be the capacity vector that is obtained from q by removing q_h. The last component $R=(R_k)_{k \in N}$ is a list of preference relations of the agents. For all $k \in N$, let P_k denote the strict relation, and I_k denote the indifference relation, associated with the preference relation R_k. We assume that, the set of agents, capacities, and preferences are common knowledge to all agents but not to the central planner.

Each intern i has a strict preference relation R_i on $\Sigma_i^H=\{\{h_1\}, ..., \{h_n\}, \varnothing\}$. Let $\mathscr{R}_i$ be the class of all such preference relations. Each hospital h has a strict preference relation R_h on $\Sigma_h^I=2^I$ which is *responsive* (Roth [5]): For all $J \subseteq I$,

1. for all $i \in I \backslash J$, $\quad J \cup \{i\}\ P_h J \Leftrightarrow \{i\}\ P_h \varnothing$;
2. for all $i, i' \in I \backslash J$, $\quad J \cup \{i\}\ P_h J \cup \{i'\} \Leftrightarrow \{i\}\ P_h \{i'\}$.

Let $\mathscr{R}_h$ be the class of all such preferences for hospital h. Let $\mathscr{R}_N=\prod_{i \in I} \mathscr{R}_i \times \prod_{h \in H} \mathscr{R}_h$. Let $\mathscr{E}(N, q)$ be the class of matching problems for a given (N, q) and $\mathscr{E}$ be the class of all matching problems. For any $R \in \mathscr{R}_N$ and $T \subset N$, let R_{-T} be the preference profile that is obtained from R by removing the preferences of the agents in T. If the group T consists of a

[3] We allow the case of 0 capacity for a hospital. This is for notational convenience: It simplifies the definition of *manipulation via pre-arranged matches.*

single agent i, then we use R_{-i} instead of $R_{-\{i\}}$. For any $R \in \mathscr{R}_N$ and $T \subset N$, let R^T be the restriction of the preference profile R to group T.

For a given pair (N, q), a *matching* μ is a mapping from N into 2^N such that:

1. $|\mu(i)| \leqslant 1$ and $\mu(i) \subseteq H$ for all $i \in I$;
2. $|\mu(h)| \leqslant q_h$ and $\mu(h) \subseteq I$ for all $h \in H$;
3. $\mu(i) = \{h\} \Leftrightarrow i \in \mu(h)$ for all $(h, i) \in H \times I$.

We denote the set of all matchings for (N, q) by $\mathscr{M}(N, q) = \mathscr{M}(H, I, q)$ and the set of all matchings by $\mathscr{M}$. Given a preference relation R_h of a hospital h, initially defined over Σ^I_h, we extend it to the set of matchings $\mathscr{M}(N, q)$, in the following natural way: h prefers the matching μ to the matching μ' if and only if it prefers $\mu(h)$ to $\mu'(h)$. We slightly abuse the notation and also use R_h to denote this extension. We perform the same extension for each intern i.

The *choice* of a hospital h from a group of interns $J \subseteq I$ under the preference R_h and capacity q_h is defined as

$$Ch_h(R_h, q_h, J) = \{J' \subseteq J : |J'| \leqslant q_h, J' R_h J'' \text{ for all } J'' \subseteq J \text{ such that } |J''| \leqslant q_h\}.$$

A matching μ is *blocked by an intern i* under (N, R, q) if $\emptyset P_i \mu(i)$. A matching μ is *blocked by a hospital h* under (N, R, q) if $\mu(h) \neq Ch_h(R_h, q_h, \mu(h))$. Note that under responsive preferences this statement is equivalent to the following: A matching μ is blocked by a hospital $h \in H$ under (N, R, q) if there is an intern $i \in \mu(h)$ such that $\emptyset P_h \{i\}$. A matching μ is *individually rational* under (N, R, q) if it is not blocked by an intern or a hospital under (N, R, q). An individually rational matching μ is *blocked by a hospital-intern pair* (h, i) under (N, R, q) if $\{h\} P_i \mu(i)$ and $\mu(h) \neq Ch_h(R_h, q_h, \mu(h) \cup \{i\})$. A matching μ is *stable* under (N, R, q) if it is not blocked by an intern, a hospital, or a hospital-intern pair. We denote the set of all stable matchings for (N, R, q) by $\mathscr{S}(N, R, q)$. Roth [4] shows that for any matching problem $(N, R, q) \in \mathscr{E}$, there exists a matching $\mu^H(N, R, q) \in \mathscr{S}(N, R, q)$ such that

$$\mu^H(N, R, q)(h) \; R_h \mu(h) \qquad \text{for all } h \in H, \text{ for all } \mu \in \mathscr{S}(N, R, q).$$

We refer to this matching as the *hospital-optimal stable matching* for (N, R, q). There is an analogous matching $\mu^I(N, R, q)$ which favors the interns and we refer to it as the *intern-optimal stable matching*.

A matching rule is a systematic procedure to select a matching for each matching problem. Formally, a *matching rule* is a function $\varphi \colon \mathscr{E} \to \mathscr{M}$ such that, for all $(N, R, q) \in \mathscr{E}$, we have $\varphi(N, R, q) \in \mathscr{M}(N, q)$. A matching rule is *stable* if it always selects a stable matching. An example of a stable matching rule is the one that selects the hospital-optimal stable matching for each problem. We denote this rule by μ^H and refer to it as the *hospital-optimal stable rule*. Another example is the rule that selects the intern-optimal stable matching for each problem. We denote it by μ^I and refer to it as the *intern-optimal stable rule*. To simplify the notation, for any matching rule φ and any agent k we use $\varphi_k(N, R, q)$ instead of $\varphi(N, R, q)(k)$.

3. MANIPULATION VIA PRE-ARRANGED MATCHES

A matching rule is manipulable via pre-arranged matches if there is a hospital and an intern who can both benefit (at least one of them strictly) by making an arrangement before the formal procedure. Formally, a matching rule φ is *manipulable via pre-arranged matches* if there is a matching problem (N, R, q) and a hospital-intern pair (h, i) such that

$$\{h\}\ R_i \varphi_i(N, R, q)$$

and

$$(\{i\} \cup \varphi_h(N \backslash \{i\}, R_{-i}^{N \backslash \{i\}}, q_{-h}, q_h - 1))\ R_h \varphi_h(N, R, q)$$

with at least one of the relations holding strictly.[4] Here, note that once the pre-arrangement is done, the intern does not participate in the formal procedure anymore while the hospital participates with one less position. There may be markets where it is illegal or undesirable for agents to match outside the centralized procedure. In such situations (as long as a stable matching rule is used) a hospital-intern pair can still pre-arrange a match as follows: The intern ranks the hospital as its top choice and the hospital ranks the intern as its top candidate. Our results have direct counterparts for this case.

3.1. *An Impossibility Result*

The United Kingdom is divided into several regional hospital-intern markets. A number of these markets have historically adopted matching rules which happen to be unstable and most of them were abandoned due to a number of problems. One of these problems is the pre-arrangement of matches between the hospitals and interns before the formal procedure. For example in Newcastle at some point 80% of the interns made such arrangements while an unstable matching rule was in use. Obviously such

[4] Postlewaite [3] introduces and studies the concept of *C-manipulation* that allows pre-arrangements between several agents in the context of exchange economies.

a problem is expected in the case of unstable matching rules.[5] What is not clear is whether stable matching rules are immune to such manipulations. Indeed pre-arranged matches are observed in U.S. medical markets as well as the Canadian layer markets in spite of the use of stable matching rules. (See Roth and Xing [11].) This empirical observation is consistent with our first result.

THEOREM 1. *There is no matching rule that is both* stable *and* non-manipulable via pre-arranged matches.[6]

Proof. Let $\varphi\colon \mathscr{E} \to \mathscr{M}$ be stable, $H=\{h_1, h_2\}$, $I=\{i_1, i_2, i_3, i_4, i_5\}$, $q_{h_1}=q_{h_2}=2$, and $R \in \mathscr{R}_N$ be such that

$$\{i_1, i_2\}\ P_{h_1}\{i_1, i_3\}\ P_{h_1}\{i_1, i_4\}\ P_{h_1}\{i_2, i_3\}$$

$$P_{h_1}\{i_1\}\ P_{h_1}\{i_2\}\ P_{h_1}\{i_3\}\ P_{h_1}\{i_4\}\ P_{h_1}\varnothing P_{h_1}\{i_5\},$$

$$\{i_2, i_3\}\ P_{h_2}\{i_2, i_5\}\ P_{h_2}\{i_1, i_3\}\ P_{h_2}\{i_1, i_4\}$$

$$P_{h_2}\{i_2\}\ P_{h_2}\{i_3\}\ P_{h_2}\{i_1\}\ P_{h_2}\{i_4\}\ P_{h_2}\{i_5\}\ P_{h_2}\varnothing,$$

$$\{h_2\}\ P_{i_1}\{h_1\}\ P_{i_1}\varnothing,$$

$$\{h_1\}\ P_{i_2}\{h_2\}\ P_{i_2}\varnothing,$$

$$\{h_1\}\ P_{i_3}\{h_2\}\ P_{i_3}\varnothing,$$

$$\{h_1\}\ P_{i_4}\{h_2\}\ P_{i_4}\varnothing,$$

$$\{h_2\}\ P_{i_5}\{h_1\}\ P_{i_5}\varnothing,$$

We have $\mathscr{S}(N, R, q)=\{\mu_1\}$, where

$$\mu_1=\begin{pmatrix} h_1 & h_2 \\ \{i_2, i_3\} & \{i_1, i_4\} \end{pmatrix}$$

and therefore $\varphi(N, R, q)=\mu_1$. Now consider an early deal between intern i_4 and hospital h_1. Intern i_4 leaves the market and hospital h_1 is left with

[5] Postlewaite [3] shows in the context of exchange economies that there is no solution that is immune to manipulation via pre-arrangements. He proves this result in two steps: In step 1 he shows that if such a solution exists at all then it should be a selection from the core correspondence. His proof of this step can be easily adopted to other domains. (Hence *manipulation via pre-arranged matches* is expected for unstable matching rules.) In step 2 he shows that there is no selection from the core correspondence that is immune to manipulation via pre-arrangements. His proof of this step is domain specific.

[6] In a related impossibility theorem, Sönmez [13] shows that there is no matching rule that is *stable* and *non-manipulable via underrepresentation of capacities*. See also Postlewaite [3], Sertel [12], and Thomson [14, 15] for applications of a similar manipulation notion in the contexts of exchange economies and public goods economies.

a capacity $q_{h_1}-1=1$. We have $\mathcal{S}(N\backslash\{i_4\}, R^{N\backslash\{i_4\}}_{-i_4}, q_{h_1}-1, q_{h_2})=\{\mu_2, \mu_3\}$, where

$$\mu_2=\begin{pmatrix} h_1 & h_2 \\ \{i_1\} & \{i_2, i_3\} \end{pmatrix}, \qquad \mu_3=\begin{pmatrix} h_1 & h_2 \\ \{i_2\} & \{i_1, i_3\} \end{pmatrix},$$

and therefore $\varphi(N\backslash\{i_4\}, R^{N\backslash\{i_4\}}_{-i_4}, q_{h_1}-1, q_{h_2})\in\{\mu_2, \mu_3\}$. We have two cases to consider:

Case 1. $\varphi(N\backslash\{i_4\}, R^{N\backslash\{i_4\}}_{-i_4}, q_{h_1}-1, q_{h_2})=\mu_2$.
For this case we have

$$\{h_1\}\ P_{i_4}\underbrace{\varphi_{i_4}(N, R, q)}_{\{h_2\}}$$

and

$$(\{i_4\}\cup\underbrace{\varphi_{h_1}(N\backslash\{i_4\}, R^{N\backslash\{i_4\}}_{-i_4}, q_{h_1}-1, q_{h_2})}_{\{i_1\}})\ P_{h_1}\underbrace{\varphi_{h_1}(N, R, q)}_{\{i_2, i_3\}}$$

and thus intern i_4 and hospital h_1 *manipulate the matching rule* φ *via pre-arranged matches.*

Case 2. $\varphi(N\backslash\{i_4\}, R^{N\backslash\{i_4\}}_{-i_4}, q_{h_1}-1, q_{h_2})=\mu_3$.
Consider the "reduced" matching problem $(N\backslash\{i_4\}, R^{N\backslash\{i_4\}}_{-i_4}, q_{h_1}-1, q_{h_2})$ and an early deal between intern i_5 and hospital h_2. Intern i_5 leaves the market and hospital h_2 is left with a capacity $q_{h_2}-1=1$. We have $\mathcal{S}(N\backslash\{i_4, i_5\}, R^{N\backslash\{i_4, i_5\}}_{-\{i_4, i_5\}}, q_{h_1}-1, q_{h_2}-1)=\{\mu_4\}$, where

$$\mu_4=\begin{pmatrix} h_1 & h_2 \\ \{i_1\} & \{i_2\} \end{pmatrix}$$

and therefore $\varphi(N\backslash\{i_4, i_5\}, R^{N\backslash\{i_4, i_5\}}_{-\{i_4, i_5\}}, q_{h_1}-1, q_{h_2}-1)=\mu_4$. Hence we have

$$\{h_2\}\ P_{i_5}\overbrace{\varphi_{i_5}(N\backslash\{i_4\}, R^{N\backslash\{i_4\}}_{-i_4}, q_{h_1}-1, q_{h_2})}^{\varnothing},$$

$$(\{i_5\}\cup\underbrace{\varphi_{h_2}(N\backslash\{i_4, i_5\}, R^{N\backslash\{i_4, i_5\}}_{-\{i_4, i_5\}}, q_{h_1}-1, q_{h_2}-1)}_{\{i_2\}})$$

$$P_{h_2}\underbrace{\varphi_{h_2}(N\backslash\{i_4\}, R^{N\backslash\{i_4\}}_{-i_4}, q_{h_1}-1, q_{h_2})}_{\{i_1, i_3\}},$$

and thus intern i_5 and hospital h_2 *manipulate the matching rule* φ *via pre-arranged matches*. This shows that φ is *manipulable via pre-arranged matches*. ▌

Recently there was an active debate in the U.S. medical community concerning the replacement of the hospital-optimal stable rule. The American Medical Student Association urged a change towards the intern-optimal stable rule. They supported their proposal with the following result.

THEOREM (Roth [5]). There is no matching rule that eliminates the incentives for hospitals to misrepresent their preferences. On the other hand, truth-telling is a dominant strategy for interns under the intern-optimal stable rule.

Based on this result, they argued that the strategic behavior on part of the interns will be eliminated under the intern-optimal stable rule. As Theorem 1 shows this is not quite correct: While the intern-optimal stable rule eliminates the incentives for interns to *unilaterally* misrepresent their preferences, an intern can still team up with a hospital to manipulate it.

3.2. *Who Can Manipulate*?

We next ask the following question: Which pairs can manipulate a stable matching rule via pre-arranged matches?

THEOREM 2. *Let* φ *be a* stable *matching rule and suppose a hospital-intern pair* (h, i) manipulates it via a pre-arranged match *under the matching problem* (N, R, q). *Then either* (i) *the intern i is one of the interns who would be assigned to hospital h in the absence of the manipulation*, *or* (ii) *the intern i is strictly worse than any intern who would be assigned to hospital h in the absence of the manipulation.*

Proof. Let $\varphi\colon \mathscr{E} \to \mathscr{M}$ be *stable* and $(N, R, q) \in \mathscr{E}$. Suppose the hospital-intern pair (h, i) *manipulates* φ *via pre-arranged matches*. That is,

$$\{h\}\ R_i \varphi_i(N, R, q)$$

and

$$(\{i\} \cup \varphi_h(N\backslash\{i\}, R_{-i}^{N\backslash\{i\}}, q_{-h}, q_h - 1))\ R_h \varphi_h(N, R, q),$$

where at least one of the relations holds strictly. Suppose $i \notin \varphi_h(N, R, q)$. Then since the preferences are strict, we have $\{h\}\ P_i \varphi_i(N, R, q)$. We want to show that

$$\{i'\}\ P_h \{i\} \qquad \text{for all} \quad i' \in \varphi_h(N, R, q).$$

Suppose on the contrary that there is an intern $i' \in \varphi_h(N, R, q)$ such that $\{i\} \; R_h\{i'\}$. Since the preferences are strict this relation holds strictly and therefore the pair (h, i) blocks the matching $\varphi(N, R, q)$ under (N, R, q), contradicting the stability of φ. ∎

That is, if there is a successful manipulation of a stable matching rule φ by a pair (h, i) via a pre-arranged match under (N, R, q), then there are two possibilities:

Type-1 *manipulation.* $i \in \varphi_h(N, R, q)$;

Type-2 *manipulation.* $\{i'\} \; P_h\{i\}$ for all $i' \in \varphi_h(N, R, q)$.

It is only the hospital that strictly benefits from a type-1 manipulation. On the other hand, both sides strictly benefit in type-2 manipulation.

REFERENCES

1. D. Gale and L. Shapley, College admissions and the stability of marriage, *Amer. Math. Monthly* **69** (1962), 9–15.
2. H. Li and S. Rosen, Unraveling in assignment markets, *Amer. Econ. Rev.* **88** (1998), 371–387.
3. A. Postlewaite, Manipulation via endowments, *Rev. Econ. Stud.* **46** (1979), 255–262.
4. A. E. Roth, The evolution of the labor market for medical interns and residents: A case study in game theory, *J. Polit. Econ.* **92** (1984), 991–1016.
5. A. E. Roth, The college admissions problem is not equivalent to the marriage problem, *J. Econ. Theory* **36** (1985), 277–288.
6. A. E. Roth, New physicians: A natural experiment in market organization, *Science* **250** (1990), 1524–1528.
7. A. E. Roth, A natural experiment in the organization of entry level labor markets: Regional markets for new physicians and surgeons in the U.K., *Amer. Econ. Rev.* **81** (1991), 415–440.
8. A. E. Roth, Proposed research program: Evaluation of changes to be considered in the NRMP algorithm, mimeo, University of Pittsburgh, 1995.
9. A. E. Roth and E. Peranson, The effects of the change in the NRMP matching algorithm, *J. Amer. Medical Assoc.* **278-9** (1997), 729–732.
10. A. E. Roth and M. Sotomayor, "Two-Sided Matching: A Study in Game Theoric Modeling and Analysis," Cambridge Univ. Press, London/New York, 1990.
11. A. E. Roth and X. Xing, Jumping the gun: Imperfections and institutions related to the timing of market transactions, *Amer. Econ. Rev.* **84** (1994), 992–1044.
12. M. Sertel, Manipulating Lindahl equilibrium via endowments, *Econ. Lett.* **46** (1994), 167–171.
13. T. Sönmez, Manipulation via capacities in two-sided matching markets, *J. Econ. Theory* **77** (1997), 197–204.
14. W. Thomson, Monotonic allocation mechanism, mimeo, University of Rochester, 1987.
15. W. Thomson, Monotonic allocation mechanisms in economies with public goods, mimeo, University of Rochester, 1987.